B3

(English Translation)
Japanese Laid-open Patent

Laid-open Number: Hei 10-97285
Laid-open Date: April 14, 1998
Application Number: Hei 08-251373
Filing Date: September 24, 1996
Applicant: Mitsubishi Electric Corporation

[Title of the Invention] Speech recognition system

[Abstract]

[Object]

To provide a speech recognition system capable of recognizing voice with a high degree of precision even when a speech with large vocabulary (large number of words) is inputted.

[Solving Means]

A use frequency score calculation section 11 is provided, which calculates use frequency scores, which assign smaller values for higher use frequency words by using information showing the use frequencies in a word-dictionary-with-frequency storage section 7 and stores the scores in a word dictionary storage section 10 and a collating section 3 obtains distance values S5 by adding, to acoustic scores showing degrees by which an inputted voice signal S1 and word dictionaries S12 are close to each other in terms of acoustics, the use frequency scores of the words stored in the word dictionary storage section 10 at a prescribed ratio. With this speech recognition system, by giving the use frequency scores, which assign smaller values for words having higher use frequencies, and adding

the scores at the time of pattern matching, it becomes possible to improve a recognition rate as to words having high use frequencies.

[Scope of Claims]

[Claim 1] A speech recognition system comprising:

an acoustic analysis section that performs acoustic analysis on an inputted voice signal at given time intervals, sequentially converts the result of the analysis into a characteristic parameter vector and voice signal power, and outputs the characteristic parameter vector and the voice signal power;

a voice section detection section that detects a voice section of the voice signal based on changes of the voice signal power received from the acoustic analysis section and outputs a voice section detection signal based on changes of a detection state of the voice section;

a collating section that performs matching between a part in the voice section of a time series of the characteristic parameter vector outputted from the acoustic analysis section and word dictionaries stored in a word dictionary storage section in accordance with an instruction from the voice section detection signal, performs pattern matching between the inputted voice signal and the word dictionaries, and outputs results of the matching as distance values;

a result output section that, when receiving an instruction from the voice section detection signal, sorts already received distance values and outputs at least one word having a small distance value as a recognition result;

a word-dictionary-with-frequency storage section in which standard patterns of recognition target words and information showing use frequencies of the words are stored; and

a use frequency score calculation section that calculates use frequency scores, which assign smaller values for higher use frequencies, by using the information showing the use frequencies in the word-dictionary-with-frequency storage section and stores the use frequency scores in the word dictionary storage section;

wherein the collating section obtains the distance values by adding the use frequency scores of the words stored in the word dictionary storage section at a prescribed ratio to acoustic scores showing degrees by which the inputted voice signal and the word dictionaries are close to each other in terms of acoustics,.

[Claim 2] A speech recognition system according to claim 1,

wherein the use frequency score calculation section is set so that the use frequency scores do not become smaller than a prescribed lower limit value.

[Claim 3] A speech recognition system according to claim 1, further comprising:

a use frequency estimation section that totals frequencies of words having the same pronunciation in an already-existing database and regards the totaled frequencies of the words with respect to the whole as a use frequency.

[Claim 4] A speech recognition system according to claim 3,

wherein when "OU" is contained in pronunciation in roman letter notation, the use frequency estimation section assumes that reading, in which "OU" is changed into "OO", is performed at a prescribed ratio, reduces a use frequency of an original word based on the prescribed ratio, adds a new word by changing "OU" into "OO", and sets a use frequency of the new word to the prescribed ratio of the use frequency of the original word before the reduction.

[Claim 5] A speech recognition system according to claim 3,

wherein when "EI" is contained in pronunciation in roman letter notation, the use frequency estimation section assumes that reading, in which "EI" is changed into "EE", is performed at a prescribed ratio, reduces a use frequency of an original word based on the prescribed ratio, adds a new word by changing "EI" into "EE", and sets a use frequency of the new word to the prescribed ratio of the use frequency of the original word before the reduction.

[Claim 6] A speech recognition system according to claim 3,

wherein the use frequency estimation section assumes that vowel sounds and nasal sounds are long-sounded at an arbitrary prescribed ratio, reduces use frequencies of original words based on the prescribed ratio, adds new words by changing the vowel sounds and the nasal sounds into long-sounded vowel sounds and long-sounded nasal sounds, and sets use frequencies of the new words to the prescribed ratio of the use frequencies of the original words before the reduction.

[Claim 7] A speech recognition system according to claim 3,

wherein the use frequency estimation section assumes that pauses are inserted between syllables at an arbitrary prescribed ratio, reduces use frequencies of original words based on the prescribed ratio, adds new words by inserting the pauses between the syllables, and sets use frequencies of the new words to the prescribed ratio of the use frequencies of the original words before the reduction.

[Claim 8] A speech recognition system according to claim 3,

wherein the use frequency estimation section assumes that double consonants are pronounced as "TSU" at an arbitrary prescribed ratio, reduces use frequencies of original words based on the prescribed ratio, adds new words by changing the double consonants into "TSU", and sets use frequencies of the new words to the prescribed ratio of the use frequencies of the original words before the reduction.

[Claim 9] A speech recognition system according to claim 3,

wherein the use frequency estimation section classifies the contents of a database in accordance with a prescribed criterion and estimates a use frequency for each classification,

the use frequency score calculation section includes a speaker recognition section that computes a use frequency score for each classification, performs speaker recognition on a voice signal of an unknown speaker using standard patterns studied from voice signals

of speakers classified in accordance with the prescribed criterion, and outputs a speaker recognition score showing to which classification the unknown speaker is close, and

the collating section obtains a matching result by adding the speaker recognition score, the use frequency scores of words in the classification, and the acoustic scores of the words at an arbitrary prescribed ratio.

[Detailed Description of the Invention]

[0001]

[Technical Field to which the Invention belongs]

The present invention relates to a speech recognition system and is applicable to recognition in which large vocabulary is set as a target word.

[0002]

[Prior Art]

In a speech recognition system in which large vocabulary, such as addresses or persons' names, are set as targets, there occurs a problem in that recognition performance deteriorates due to the existence of many analogous words and a computation amount becomes enormous because pattern matching with large vocabulary is performed, which makes it extremely difficult to realize such a speech recognition system. A conventional speech recognition system of this type, in which large vocabulary is set as a target, disclosed in JP 03-84600 A is shown in FIG. 9.

[0003]

An acoustic analysis section 1 performs acoustic analysis on an inputted voice signal S1 at given time intervals, converts the result of the analysis into a characteristic parameter vector S2 and into voice signal power S3, and outputs the vector S2 and the power S3. A voice section detection section 2 detects a voice section of the voice signal based on changes of the voice signal power S3 received from the acoustic analysis section 1 and outputs a voice section detection signal S4 based on changes of the detection state of the voice section. A collating section 3 performs matching between a part in the voice section of a time series of the characteristic parameter vector S2 received from the acoustic analysis section 1 and word dictionaries S6 in the order of reading from a word dictionary storage section 5 in accordance with an instruction from the voice section detection signal S4 and sequentially outputs acoustic scores indicating degrees, by which the inputted voice signal S1 and the word dictionaries S6 are close to each other in terms of acoustics, as distance values S5.

[0004]

It should be noted here that a word-dictionary-with-frequency storage section 7 stores labels showing readings of recognition target words and information showing use frequencies, a word dictionary sorting section 6 sorts word information in a decreasing order of the use frequencies in the word-dictionary-with-frequency

storage section 7, and the word dictionary storage section 5 stores the sorted word information. Also, in the drawing, S6 denotes word dictionaries, S7 word dictionaries, and S8 word dictionaries with frequencies. When receiving an input from the voice section detection signal S4 or receiving an output request signal S9 inputted from the outside, a result output section 4 sorts distance values S5, which are already received but are not yet outputted, based on the distance values S5 and outputs one word or multiple words, whose distance values S5 are small, as a recognition result S10. [0005]

An operation of the speech recognition system having such a construction will be described. Prior to recognition, the word dictionary sorting section 6 reads out the contents of the word-dictionary-with-frequency storage section 7, performs sorting based on use frequencies, and accumulates the outcome in the word dictionary storage section 5 in a decreasing order of the use frequencies. Hereinafter, an operation at the time of recognition will be described. At the recognition apparatus, processing is performed in unit times of around 10 milliseconds. The unit times will be referred to as the "frames". The acoustic analysis section 1 repeats an operation, in which it performs acoustic analysis on the inputted voice signal S1 and converts the result of the analysis into the characteristic parameter vector S2 and the voice signal power S3 for each frame. As an acoustic analysis technique, a

technique based on LPC (Linear Prediction coefficient) analysis, FFT (fast Fourier transform), or filter bank is used, for instance.

[0006]

Next, an operation of the voice section detection section 2 will be described. The voice section detection section 2 performs voice section detection in which it monitors the voice signal power S3, detects a start point of a voice section when the voice signal power S3 exceeds a certain threshold value, detects an end point candidate of the voice section when the voice signal power S3 falls below the threshold value, and judges that the end point candidate is correct and makes end point determination when the state, in which the voice signal power S3 falls below the threshold value, continues for a certain time. This time is generally set at around 0.3 second. When the voice signal power S3 rises and exceeds the threshold value again before 0.3 second has passed, the voice section detection section 2 invalidates the end point candidate.

[0007]

The operation of the voice section detection section 2 will be described based on a concrete example based on FIG. 10. Changes of the voice signal power in the case where "HOTTA" is vocalized are shown in FIG. 10 as an example. The horizontal axis represents time and the vertical axis indicates the magnitude of the voice signal power. It is assumed that a section from a frame T1 to a frame T2 is a vocalization section of "HO", a section from the frame

T2 to a frame T3 is a vocalization section of "T", and a section from the frame T3 to a frame T4 is a vocalization section of "TA". In FIG. 10, the voice signal power rises from a noise level and exceeds a threshold value P1 at a point in time of the frame T1 and falls therebelow in the frame T2. Then, the voice signal power exceeds the threshold value P1 in the frame T3 and falls therebelow in the frame T4 again. A frame T5 indicates a point in time when 0.3 second has passed from the frame T4. The first "T" of "HOTTA" is acoustically classified into a double consonant. In ordinary vocalization, such a double consonant has a time length of 0.3 second or less. Therefore, also in this example, a time between the frame T2 and the frame T3 is set at 0.3 second or less. According to the operation of the voice section detection section 2 described above, a section from the frame T1 to the frame T4 is detected as a voice section.

[0008]

The voice section detection section 2 sends out three kinds of signals that are a start point signal, an end point candidate signal, and an end point determination signal as the voice section detection signal S4. In FIG. 10, the voice section detection section 2 sends out the start point signal in the frame T1 and the frame T3, sends out the end point candidate signal in the frame T2 and the frame T4, and sends out the end point determination signal in the frame T5. When the end point determination signal is not sent

out but the start point signal is sent out after the end point candidate signal, this indicates that an end point candidate designated by the end point candidate signal before the start point signal, that is, the frame T2 should be invalidated.

[0009]

The collating section 3 receives the characteristic parameter vector S2 in the voice section sent from the acoustic analysis section 1 and internally accumulates it in a section from the start point signal to the end point determination signal of the voice section detection signal S4. When receiving the end point candidate signal from the voice section detection section 2 as the voice section detection signal S4, the collating section 3 starts pattern matching. In FIG. 10, each frame, in which pattern matching is performed, is indicated with diagonal shading. There are various methods of the pattern matching and a method based on DP (Dynamic Programming) matching or HMM (Hidden Markov Model) is applicable, for instance. The collating section 3 reads out the word dictionaries S6 in the word dictionary storage section 5 in the order of arrangement, performs pattern matching against the internally accumulated characteristic parameter vector S2 in the section from the frame T1 to the frame T2, and sends out a distance value S5 to the result output section 4. The word dictionaries S6 are arranged in the word dictionary storage section 5 in a decreasing order of frequencies, so the pattern matching is performed so that a word having the highest

frequency is processed first.

[0010]

Following this, when receiving a start point signal from the voice section detection section 2 as the voice section detection signal S4 in the frame T3, the collating section 3 terminates the pattern matching by regarding the section from the frame T1 to the frame T2 as invalid. Further, following this, when receiving an end point candidate signal from the voice section detection section 2 as the voice section detection signal S4 in the frame T4, the collating section 3 performs pattern matching against the internally accumulated characteristic parameter vector S2 in the section from the frame T1 to the frame T4 and sequentially outputs acoustic scores indicating degrees, by which the inputted voice signal S1 and the word dictionaries S6 are close to each other, and the words to the result output section 4 as the distance values S5.

[0011]

The result output section 4 sequentially performs sorting on the distance values S5 sent from the collating section 3 based on the acoustic scores. When receiving a start point signal as the voice section detection signal S4 from the voice section detection section 2, the result output section 4 clears the distance values S5 sorted by that time. When receiving an end point determination signal as the voice section detection signal S4 from the voice section detection section 2, the result output section 4 outputs one result

or multiple recognition results S10 ranked high among the distance values S5 sorted by that time. When a vocalization person confirms a screen, on which the outputted results are displayed, and judges that a correct recognition result is not contained, he/she inputs an output request signal S9. When the output request signal S9 is inputted, the result output section 4 outputs one word or multiple words ranked high among the distance values S5, which are already sorted but are not yet outputted at that point in time, as the recognition result S10. When a recognition result is not contained therein, the sequence described above is further repeated.

[0012]

A flow of the processing by the result output section 4 will be further described with reference to FIG. 10. In the frame T1, the result output section 4 clears internal data. From the frame T2, distance values S5 are transferred from the collating section 3, so the result output section 4 sequentially performs sorting based on the distance values S5 and internally accumulates the result of the sorting. In the frame T3, the result output section 4 clears the sorting result. In the frame T4, distance values S4 are transferred from the collating section 3 again, so the result output section 4 sequentially performs sorting based on the acoustic scores and internally accumulates the result of the sorting. In the frame T5, the result output section 4 outputs one word or multiple words ranked high among the sorted distance values S5 as a recognition

result S10. In FIG. 10, each period, during which a recognition result S10 is outputted, is indicated with a rectangle filled in with black. In a like manner, after the frame T5, distance values S5 are transferred from the collating section 3 and the result output section 4 sequentially performs sorting and internally accumulates the result of the sorting. In the frame T6, an output request signal S9 is inputted from the outside, so the result output section 4 outputs one word or multiple words ranked high among the distance values S5 sorted by that point in time as a recognition result S10.

[0013]

As described above, with the speech recognition system according to the conventional technique, when word recognition of large vocabulary is performed, in the case of a word having a high frequency, a recognition result is outputted when 0.3 second has passed after vocalization is ended. Also, even in the case of a word having a low frequency, it is possible to obtain a recognition result by sending an output request signal S9 to the apparatus after a while.

[0014]

[Problems to be solved by the Invention]

However, the conventional speech recognition system is constructed in the manner described above, so it is impossible to recognize words having low frequencies before 0.3 second has passed after vocalization is ended no matter how carefully the words are

vocalized. Also, when it is attempted to recognize large vocabulary, such as persons' names, including several ten thousand words, the number of analogous words, such as "Oono/Ono" and "Satou/Sato", is increased, so there occurs a problem in that a recognition rate is lowered. FIG. 11 shows recognition performance of the conventional speech recognition system with respect to voice data concerning vocalization of person's names collected through a telephone line. In the drawing, the vertical axis represents an error rate and the horizontal axis indicates the number of words having high frequencies used in pattern matching in a form of logarithm (log). In the drawing, the solid line represents the error rate of the conventional speech recognition system and the dotted line indicates the rate of a situation in which no correct word is contained in words in the word dictionary storage section 5 used in pattern matching.

[0015]

The total number of Japanese person's names is around 58,000 and the number of words, with which pattern matching is performed, is increased in a rightward direction on the drawing. As can be seen from the drawing, when recognition is performed using 1,000 words having high frequencies, an omission rate that is a rate, at which a correct word is not contained in the 1,000 words, becomes 30.7%. In addition, there exists a misrecognition rate that is 16.5%, so an incorrect rate becomes 47.2% of vocalization in total. In

contrast to this, when matching is performed against 57,711 words using a longer time for the matching, the incorrect rate becomes 63.4% because the omission rate becomes 1.3% and the misrecognition rate becomes 62.1%. That is, when the number of words is increased, the recognition rate is extremely lowered, which results in a problem in that the incorrect rate is increased.

[0016]

The present invention has been made in order to solve the problems described above and provides a speech recognition system that is capable of recognizing voice with a high degree of precision even in the case of large vocabulary.

[0017]

[Means for solving the problems]

A speech recognition system according to the present invention includes: an acoustic analysis section that performs acoustic analysis on an inputted voice signal at given time intervals, sequentially converts the result of the analysis into a characteristic parameter vector and voice signal power, and outputs the characteristic parameter vector and the voice signal power; a voice section detection section that detects a voice section of the voice signal based on changes of the voice signal power received from the acoustic analysis section and outputs a voice section detection signal based on changes of a detection state of the voice section; a collating section that performs matching between a part

in the voice section of a time series of the characteristic parameter vector outputted from the acoustic analysis section and word dictionaries stored in a word dictionary storage section in accordance with an instruction from the voice section detection signal, performs pattern matching between the inputted voice signal and the word dictionaries, and outputs results of the matching as distance values; a result output section that, when receiving an instruction from the voice section detection signal, sorts already received distance values and outputs at least one word having a small distance value as a recognition result; a word-dictionary-with-frequency storage section in which standard patterns of recognition target words and information showing use frequencies of the words are stored; and a use frequency score calculation section that calculates use frequency scores, which assign smaller values for higher use frequencies, from the information showing the use frequencies in the word-dictionary-with-frequency storage section and stores the use frequency scores in the word dictionary storage section; wherein the collating section obtains the distance values by adding, to acoustic scores showing degrees by which the inputted voice signal and the word dictionaries are close to each other in terms of acoustics, the use frequency scores of the words stored in the word dictionary storage section at a prescribed ratio.

[0018]

Further, in the speech recognition system according to another aspect of the present invention, the use frequency score calculation section is set so that the use frequency scores do not become lower than a prescribed lower limit value.

[0019]

Further, the speech recognition system according to another aspect of the present invention further includes: a use frequency estimation section that totals frequencies of words having the same pronunciation in an already-existing database and regards the totaled frequencies of the words with respect to the whole as a use frequency.

[0020]

Further, in the speech recognition system according to another aspect of the present invention, wherein when "OU" is contained in pronunciation in roman letter notation, the use frequency estimation section assumes that reading, in which "OU" is changed into "OO", is performed at a prescribed ratio, reduces a use frequency of an original word based on the prescribed ratio, adds a new word by changing "OU" into "OO", and sets a use frequency of the new word to the prescribed ratio of the use frequency of the original word before the reduction.

[0021]

Further, in the speech recognition system according to another aspect of the present invention, when "EI" is contained in

pronunciation in roman letter notation, the use frequency estimation section assumes that reading, in which "EI" is changed into "EE", is performed at a prescribed ratio, reduces a use frequency of an original word based on the prescribed ratio, adds a new word by changing "EI" into "EE", and sets a use frequency of the new word to the prescribed ratio of the use frequency of the original word before the reduction.

[0022]

Further, in the speech recognition system according to another aspect of the present invention, the use frequency estimation section assumes that vowel sounds and a nasal sounds are long-sounded at an arbitrary prescribed ratio, reduces use frequencies of original words based on the prescribed ratio, adds new words by changing the vowel sounds and the nasal sounds into long-sounded vowel sounds and long-sounded nasal sounds, and sets use frequencies of the new words to the prescribed ratio of the use frequencies of the original words before the reduction.

[0023]

Further, in the speech recognition system according to another aspect of the present invention, the use frequency estimation section assumes that pauses are inserted between syllables at an arbitrary prescribed ratio, reduces use frequencies of original words based on the prescribed ratio, adds new words by inserting the pauses between the syllables, and sets use frequencies of the new words

to the prescribed ratio of the use frequencies of the original words before the reduction.

[0024]

Further, in the speech recognition system according to another aspect of the present invention, the use frequency estimation section assumes that double consonants are pronounced as "TSU" at an arbitrary prescribed ratio, reduces use frequencies of original words based on the prescribed ratio, adds new words by changing the double consonants into "TSU", and sets use frequencies of the new words to the prescribed ratio of the use frequencies of the original words before the reduction.

[0025]

Further, in the speech recognition system according to another aspect of the present invention, the use frequency estimation section classifies the contents of a database in accordance with a prescribed criterion and estimates a use frequency for each classification, the use frequency score calculation section includes a speaker recognition section that computes a use frequency score for each classification, performs speaker recognition on a voice signal of an unknown speaker using standard patterns studied from voice signals of speakers classified in accordance with the prescribed criterion, and outputs a speaker recognition score showing to which classification the unknown speaker is close, and the collating section obtains a matching result by adding the speaker recognition

score, the use frequency scores of words in the classification, and the acoustic scores of the words at an arbitrary prescribed ratio.

[0026]

[Embodiment Mode of the Invention]

Hereinafter, embodiments of the present invention will be described with reference to the accompanying drawings.

[0027] First Embodiment

In FIG. 1 in which each portion corresponding to a portion in FIG. 9 is given the same reference numeral, a speech recognition system according to a first embodiment of the present invention is shown. Like in the case of the conventional speech recognition system described above with reference to FIG. 9, an acoustic analysis section 1 performs acoustic analysis on an inputted voice signal S1 at given time intervals, converts the result of the analysis into a characteristic parameter vector S2 and voice signal power S3, and outputs the vector S2 and the power S3. A voice section detection section 2 detects each voice section of the voice signal S1 based on changes of the voice signal power S3 received from the acoustic analysis section 1 and outputs a voice section detection signal S4 based on changes of the detection state of the voice section.

[0028]

A collating section 3 performs matching between a part in the voice section of a time series of the characteristic parameter vector

received from the acoustic analysis section 1 and word dictionaries with scores S12 in the order of reading from a word dictionary storage section 10 in accordance with an instruction from the voice section detection signal S4, adds acoustic scores indicating degrees, by which the inputted voice signal S1 and the word dictionaries with scores S12 are close to each other, and use frequency scores to each other at a prescribed ratio, and sequentially outputs addition results as distance values S5. Here, in this first embodiment, a word-dictionary-with-frequency storage section 7 stores labels showing readings of recognition target words and information showing use frequencies and a use frequency score calculation section 11 adds use frequency scores to word dictionaries with frequencies S8 in accordance with the information showing the use frequencies in the word-dictionary-with-frequency storage section 7 and outputs addition results as word dictionaries with scores S11 in a decreasing order of the use frequencies. The word dictionary storage section 10 stores the word dictionaries with scores S11 in a decreasing order of the use frequencies.

[0029]

A result output section 4 receives input of the voice section detection signal S4 or an output request signal S9 inputted from the outside, sorts distance values S5 that are already received but are not yet outputted, and outputs one word or multiple words having small distance values S5 as a recognition result S10.

[0030]

An operation of the speech recognition system having such a construction will be described. Prior to recognition, the use frequency score calculation section 11 reads out the contents of the word-dictionary-with-frequency storage section 7, obtains the word dictionaries with scores S11 from the use frequencies, and stores them into the word dictionary storage section 10 in a decreasing order of the use frequencies. As a method of giving the use frequency scores, for instance, there is a method with which the use frequency scores are obtained using the following arithmetic expression.

[0031]

[Expression 1]

$$S(w) = -1.0 \times \log(P(w)) \cdots\cdots (1)$$

[0032]

In Expression (1), w is a word, P(w) is the use frequency of the word w expressed in a form of probability, and S(w) is the use frequency score of the word w. Here, S(w) assumes a small value in the case of a word having a high use frequency and assumes a great value in the case of a word having a large use frequency. However, when the use frequency is too small, the use frequency score becomes an extremely great value and outputting as a recognition result ranked at a high place at the result output unit 4 becomes impossible no matter how carefully vocalization is performed, so a construction may be used in which by providing a lower limit value

for the use frequency scores, it is made possible to output even a word having an extremely low use frequency as a recognition result when its acoustic score is small.

[0033]

An operation of this speech recognition system at the time of recognition will be described. Operations of the acoustic analysis section 1, the voice section detection section 2, and the result output section 4 are the same as those of the conventional speech recognition system described with reference to FIGS. 9 to 11, so the description thereof will be omitted in this embodiment. In the following description, an operation of the collating section 3 that is a feature of this first embodiment will be explained. The collating section 3 sequentially reads out the word dictionaries with scores S12 in the word dictionary storage section 10 and performs pattern matching like in the case of the conventional speech recognition system. In this embodiment, the collating section 3 also adds a use frequency score S(w) to an acoustic score D (w) by means of a weight R as expressed by the following expression.

[0034]

[Expression 2]

$$D'(w)=D(w)+R\times S(w)\cdots\cdots(2)$$

[0035]

As a result, each word having a low use frequency score is made easy to be recognized and each word having a high score is

made difficult to be recognized. That is, an effect is given by which each word having a high use frequency is made easy to be recognized and each word having a low use frequency is made difficult to be recognized. A recognition experiment was conducted according to this first embodiment under the same condition as in the case of the recognition experiment described above with reference to FIG. 11 and it was found that the error rate in the case of the 57,711 recognition target words was improved from 63.4% to 32.1%.

[0036] Second Embodiment

In the first embodiment described above, explanation has been made by assuming that the use frequency score calculation section 11 has a function of performing sorting in a decreasing order of the use frequencies and performing storing into the word dictionary storage section 10, although when H/W is sufficiently fast and it is possible to perform pattern matching against every word candidate at high speed or when it is possible to obtain a matching result for every candidate through pattern matching before the frame T5 in FIG. 10 by additionally adopting a computation-amount-reduction measure represented by a beam search method or a pruning method described in an article entitled "Comparison of Total Search Method·Beam Search Method·A*Search Method in Isolated-Word Speech Recognition" (collected lecture papers in 1996 Spring Meeting of the Acoustical Society of Japan, 2-5-10, written by Masaki Ida and Seiichi Nakagawa), the necessity for division at the time of pattern

matching in the case of the conventional speech recognition system is eliminated and, in addition, the function of the use frequency score calculation section 11 of performing sorting in a decreasing order of the use frequencies and performing storing into the word dictionary storage section 7 becomes unnecessary.

[0037]

A speech recognition system having a collating section 3 that is capable of performing such high-speed pattern matching will be described in this second embodiment. The construction of this speech recognition system is the same as that in the first embodiment, so the description thereof will be omitted in this embodiment. An operation of the speech recognition system having such a construction will be described. Prior to recognition, the use frequency score calculation section 11 reads out the contents of the word-dictionary-with-frequency storage section 7, obtains word dictionaries with scores S11 from use frequencies, and stores them in the word dictionary storage section 10. In the word dictionary storage section 10, arrangement in a decreasing order of the use frequencies is not required and random arrangement is possible. In order to give use frequency scores, the same method as in the first embodiment described above is used.

[0038]

Operations of the acoustic analysis section 1 and the voice section detection section 2 are the same as those of the conventional

speech recognition system described above with reference to FIGS. 9 to 11, so the description thereof will be omitted in this embodiment. FIG. 2 is a timing chart for an explanation of an operation of the speech recognition system according to this second embodiment. Hereinafter, operations of the collating section 3 and the result output section 4 will be described with reference to FIG. 2. Processing in frames before the frame T5 is the same as that in the case of the conventional speech recognition system. The collating section 3 according to this second embodiment has a sufficiently high processing capability, so pattern matching processing is ended before the frame T5. Therefore, in the frame T5, the result output section 4 sorts distance values S5 transferred from the collating section 3 and outputs one word or multiple words having small composite scores as a recognition result S10 in response to an end point determination signal that is the voice section detection signal S4 from the voice section detection section 2. Also, when receiving an output request signal S9 from the outside, the result output section 4 outputs one word or multiple words having small distance values S5 as a recognition result S10 in addition to the already outputted recognition result S10.

[0039] Third Embodiment

In the first embodiment and the second embodiment described above, a speech recognition system has been described which adopts a system in which pattern matching is performed for one word at

a time after a word end point candidate is determined, although it is possible to provide the same effect even by using the collating section 3 that performs frame-synchronization-type pattern matching. The frame-synchronization-type pattern matching is a method with which pattern matching with respect to every word dictionary is carried out at the same time and has a shortcoming that a work memory amount is greatly increased as compared with the method with which the pattern matching is performed for one word at a time, but has such a feature that it is made possible to perform pattern matching with efficiency because it is possible to perform the pattern matching in parallel with voice input. The frame-synchronization-type pattern matching uses a method described in an article entitled "Acceleration of DP Matching through Integration of Frame Synchronization, Beam Search, and Vector Quantization" (Journal D of the Institute of Electronics, Information and Communication Engineers, Vol. J71-D, No. 9, pp 1650-1659, written jointly by Hiroaki Sakoe, Hiromi Fujii, Kazunaga Yoshida, and Nobuo Watari), for instance.

[0040]

The construction of such a speech recognition system is the same as the construction in the first embodiment, so the description thereof will be omitted in this embodiment. An operation of the speech recognition system according to this third embodiment will be described. Operations of the acoustic analysis section 1 and

the voice section detection section 2 are the same as those in the second embodiment, so the description thereof will be described in this embodiment. Operations of the collating section 3 and the result output section 4 will be described with reference to FIG. 3. First, the operation of the collating section 3 will be described. The collating section 3 starts pattern matching processing in response to a start point signal that is one kind of the voice section detection signal S4 from the voice section detection section 2 and ends the operation in response to an end point determination signal that is one kind of the voice section detection signal S4.

[0041]

Distance values S5 are outputted from the collating section 3 in each frame. The result output section 4 sorts the distance values S5 in an end point candidate frame in response to an end point candidate signal that is one kind of the voice section detection signal S4 and outputs one word or multiple words having small distance values 11 as a recognition result S10 in response to an end point determination signal that is one kind of the voice section detection signal S4. In FIG. 3, there are two end point candidate signals in frame T2 and T4, although a recognition result outputted in the frame T5 is a recognition result S10 obtained in the frame T4. By performing the frame-synchronization-type pattern matching in this manner, it becomes possible to perform the processing by the collating section 3 also in a section from the frame T1 to the frame T2 and

a section from the frame T3 to the frame T4 in which computation is not performed in the case of the conventional speech recognition system, which makes it possible to realize an efficient speech recognition system.

[0042] Fourth Embodiment

In the above description, a case has been described in which the use frequencies of words are known in advance. Here, there is a case where it is possible to obtain the use frequencies by operating the speech recognition system for a while, although it is difficult to obtain the use frequencies at an initial stage of the operation in many cases. In a residents' ledger existing at a self-governing body or a customer database or a personnel database owned by a company, however, addresses, persons' names, telephone numbers, genders, ages, and the like are recorded, for instance. Therefore, in an information service system for residents or the like, for instance, it is possible to estimate the use frequencies of words from the contents of the residents' ledger. That is, words, such as addresses, persons' names, or telephone numbers, which correspond to high population ratios are estimated to have high use frequencies. The same estimation is possible also in the case of the customer database or the personnel database owned by a company. In the fourth embodiment, a speech recognition system that estimates use frequencies will be described by taking, as an example, a case where the use frequencies of persons' names are estimated.

[0043]

The construction of a speech recognition system according to this fourth embodiment is shown in FIG. 4 by giving each portion corresponding to a portion in FIG. 1 the same reference numeral. In FIG. 4, an acoustic analysis section 1, a voice section detection section 2, a collating section 3, a result output section 4, a word-dictionary-with-frequency storage section 7, a use frequency score calculation section 11, and a word dictionary storage section 10 are the same as those in the third embodiment, so the description thereof will be omitted in this embodiment. In the drawing, a database 12 is a database, which contains the person's names of residents, and it is assumed that the pronunciation of the person's names is indicated using kana letters in the database. Also, a use frequency estimation section 13 is a section that generates name frequency information and readings from the database 12. Further, S13 denotes persons' name information and S14 indicates word dictionaries with frequencies.

[0044]

An estimation method used by the word-dictionary-with-frequency storage section 7 will be described. First, the word-dictionary-with-frequency storage section 7 searches the database 12 and finds population N (w) for each word w by setting person's names having the same pronunciation as one word w. Counting is performed by regarding words having different

Chinese characters as the same word w when they are pronounced in the same manner. Following this, when the number of persons having a person's name is large, it is estimated that the use frequency of the person's name is high and a use frequency P(w) is obtained using the following expression.

[0045]

[Expression 3]

$$P(w)=N(w)/(ALLN)\cdots\cdots(3)$$

[0046]

In Expression (3), ALLN is the total number of persons contained in the database 12. As the readings of words, the readings contained in the database 12 are used. The subsequent method of creating word dictionaries with scores S11 and operations of the acoustic analysis section 1, the voice section detection section 2, the collating section 3, and the result output section 4 are the same as those in the third embodiment, so the description thereof will be omitted in this embodiment.

[0047] Fifth Embodiment

In the fourth embodiment described above, an example has been described in which the pronunciation contained in the database 12 is used as the readings of the words in the word-dictionary-with-frequency storage section 7. Generally, however, the pronunciation contained in the database 12 is given in kana letter notation for writing, which leads to a case where

the pronunciation does not conform with vocalization inputted into the speech recognition system. For instance, a person's name in the database 12 that is "SATOU" in the kana letter notation for writing is vocalized as "SATŌ" using a long sound with a probability of 80%. With the remaining probability of 15%, the person's name is vocalized as "SATOU" in the kana letter notation for writing. The person's name is also vocalized as "SA, TO, U" by inserting pauses between kana letters. All of these vocalization should be voice-recognized as designating the same person's name, so by automatically adding these words and estimating their use frequencies, it becomes possible to improve the recognition rate. [0048]

The construction of a speech recognition system according to this fifth embodiment is the same as that according to the fourth embodiment shown in FIG. 4, so the description thereof will be omitted in this embodiment. However, an operation of the use frequency estimation section 13 in FIG. 4 is changed from that described in the fourth embodiment and the use frequency estimation section 13 is given a new function of adding words for different readings with respect to one pronunciation using a word transformational rule. Hereinafter, an operation of the use frequency estimation section 13 in this fifth embodiment will be described. FIG. 5 is a flowchart showing an operation of the use frequency estimation section 13 in the present invention. In the drawing, the processing starts

at "START" and ends at "END". First, in step ST1 in the drawing, it is judged whether a word contains "OU" in roman letter notation. When the result of this judgment is positive, the processing proceeds to step ST2 in which a word is added by changing "OU" into "OO". The use frequency of the added word is obtained by multiplying the use frequency of the original word by 0.8. Then, the use frequency of the original word containing "OU" is multiplied by 0.2.

[0049]

Next, in step ST3, it is judged whether a word contains "EI" in roman letter notation. When the result of this judgment is positive, the processing proceeds to step ST4 in which a word is added by changing "EI" into "EE". The use frequency of the added word is obtained by multiplying the use frequency of the original word by 0.7. Then, the use frequency of the original word containing "EI" is multiplied by 0.3. Following this, in step ST5, it is judged whether a word contains a double consonant. When the result of this judgment is positive, there are some persons who vocalize the double consonant as "TSU", so the processing proceeds to step ST6 in which a word is added by changing the double consonant into "TSU". The use frequency of the added word is obtained by multiplying the use frequency of the original word by 0.05. Then, the use frequency of the original word containing the double consonant is multiplied by 0.95.

[0050]

Next, in step ST7, a long-sounded word and a paused word are added for every word. The use frequency of the long-sounded word is obtained by multiplying the use frequency of the original word by 0.1 and the use frequency of the paused word is obtained by multiplying the use frequency of the original word by 0.05. Then, the use frequency of the original word is changed through multiplication by 0.85. Here, in long-sounding of the vowel sounds and the nasal sounds, there is also a case where last syllables are not long-sounded, so such a transformational rule may be used.

[0051]

A concrete result of the processing by the speech recognition system having such a construction will be described below. Here, it is assumed that the use frequency estimation section 13 described in the fourth embodiment estimates words and use frequencies shown in FIG. 6 in the case where a certain database 12 is used. In contrast to this, the use frequency estimation section 13 in this fifth embodiment estimates 20 words shown in FIG. 7. In FIG. 7, each hyphen (-) indicates that a vowel sound or the nasal sound is long-sounded and each dot (·) indicates that a pause is inserted between syllables. A transformational rule is used under which in the long-sounding of the vowel sounds and the nasal sound, last syllables is not long-sounded.

[0052]

The use frequency of each word is obtained through

multiplication by a prescribed numerical value in accordance with the flow shown in FIG. 5. For instance, when the rule in step ST7 shown in FIG. 5 is applied to "ABE", "A-BE" is added as the result of the long-sounding of the vowel sound and "A·BE" is added as a result of the insertion of a pause between the syllables. The use frequency of "ABE" is originally 0.04598 and is changed through multiplication by 0.85. Also, the use frequency of "A-BE" is obtained through multiplication by 0.10 and the use frequency of "A·BE" is obtained through multiplication by 0.05. Here, as to "NITTA", a word obtained as a result of the insertion of a pause between syllables becomes the same as the original word, so the use frequency of "NITTA" is multiplied by 0.90.

[0053]

As described above, according to the fifth embodiment, various vocalization forms and their use frequencies are estimated from the pronunciation in the database 12, so it becomes possible to realize a speech recognition system exhibiting favorable recognition performance. Note that the values for the multiplication described above are empirically obtained from arbitrary search results, although they may be changed in accordance with a database.

[0054] Sixth Embodiment

There is a case where population is biased in the contents of a database. For instance, there exist differences in population between male names and female names. Therefore, when information

showing the gender of a voice signal is added, it becomes possible to further enhance the recognition performance. A construction of a speech recognition system according to this sixth embodiment is shown in FIG. 8. In the drawing, constructions of an acoustic analysis section 1, a voice section detection section 2, a collating section 3, a result output section 4, a word-dictionary-with-frequency storage section 7, a use frequency score calculation section 11, a word dictionary storage section 10, a database 12, and a use frequency estimation section 13 are the same as those in the fifth embodiment described above.

[0055]

In FIG. 8, a speaker recognition section 14 is a section that performs speaker recognition through comparison of a characteristic parameter vector S2 from the acoustic analysis section 1 in accordance with a voice section detection signal S4 from the voice section detection section 2. In the following description, a case where the genders of speakers are recognized and first names of person's names are classified into male names and female names and are stored will be explained as an example.

[0056]

First, prior to recognition, the use frequency estimation section 13 creates word dictionaries with frequencies S14 by generating a male word and a female word for the same name in the database 12. Then, the use frequency score calculation section 11

calculates a score for each of the male word and the female word and stores it in the word dictionary storage section 10. As a result, the storage amounts of the word-dictionary-with-frequency storage section 7 and the word dictionary storage section 10 are doubled. Also, in the speaker recognition section 14, standard patterns for speaker recognition are stored. As a method of the speaker recognition, many methods are proposed, although a method using vector quantization will be described as an example in this embodiment.

[0057]

In the speaker recognition section 14, M standard patterns 1 are prepared for males and M standard patterns 2 are prepared for females. The standard patterns are respectively studied from male voice signals and female voice signals using an LBG (Linde Buzo Gray) algorithm or the like. When the mth standard pattern for a gender i is referred to as M(i, m) and the characteristic parameter vector 9 in a frame t is referred to as "L(t)", it is possible to obtain S2(i) that is a speaker recognition score 27 using the following arithmetic expression.

[0058]

[Expression 4]

$$S2(i)=\frac{\sum_{t=T1}^{T4}\min(dis(M(i,m),L(t)))_{m=1,M}}{T4-T1+1} \cdots\cdots (4)$$

[0059] [0060] [0061]

In Expression (4), the following expression means the minimum value as to "m=1, M" of the element "X(m)".

[Expression 5]

$$\min(X(m))_{m=1,\ M} \cdots\cdots (5)$$

[0062] [0063]

Also, the following expression means a distance value between M(i, m) and L(T).

[Expression 6]

$$dis(M(i,\ m),\ L(t)) \cdots\cdots (6)$$

It is possible to collectively perform computation expressed by Expression (4) in the frame T4 and it is also possible to perform the computation in a frame synchronization manner from the frame T1. The frames T1 and T4 are informed as a voice section detection signal S4. The speaker recognition score S15 obtained in this manner is added to an acoustic score and a use frequency score at a ratio of R2 at the collating section 3 and smaller one of a result for male and a result for female is set as a final matching result.

[0064]

[Expression 7]

$$D'(w)=\min(D(w)+R\times S1(i,\ w)+R2\times S2(i))_{i=1,\ 2} \cdots\cdots (7)$$

[0065]

In Expression (7), D(w) and R are the same those used in Expression (2) and S1(i, w) is a use frequency score with respect to a word w for a gender i.

[0066]

In the above description, the standard patterns of the speaker recognition section 14 have been explained as standard patterns that are different from the standard patterns of the word dictionary storage section 10, although in the case of a multi-template speech recognition system in which the word dictionary storage section 10 stores standard patterns for males and females, it is possible to divert the standard patterns to the speaker recognition. Even with this construction, it is possible to provide the same effect as above. Also, in the above description, an example has been explained in which the speaker recognition is performed based on genders, although it is also possible to perform the speaker recognition by classifying the database 12 based on ages, the languages of names examples of which are Japanese and English, or the like. Even in this case, it is possible to provide the same effect.

[0067]

[Effects of the Invention]

As described above, according to the present invention, distance values are obtained by adding use frequency scores calculated from the use frequencies of words to acoustic scores at a prescribed ratio, so it becomes possible to enhance recognition performance with respect to words having high frequencies, which makes it possible to realize a speech recognition system that is

capable of significantly improving a recognition rate as a whole even in the case of large vocabulary.

[0068]

Also, according to the present invention, a lower limit value is set for the scores of words having extremely low use frequencies, so even in the case of such words having extremely low use frequencies, when acoustic scores are favorable, it becomes possible to obtain recognition results ranked high, which makes it possible to realize a speech recognition system that is capable of significantly improving a recognition rate as a whole even in the case of large vocabulary.

[0069]

Further, according to the present invention, it is possible to estimate use frequencies from an already-existing database, so it becomes possible to set use frequency scores even for words whose use frequencies are unclear, which makes it possible to realize a speech recognition system that is capable of significantly improving a recognition rate as a whole even in the case of large vocabulary.

[0070]

Still further, according to the present invention, when there exist words whose pronunciation given by an already-existing database contains "OU" in roman letter notation, words are added by changing "OU" into "OO" and their use frequencies are set based

on a prescribed ratio, so it becomes possible to recognize the words even when they are vocalized in manners different from the pronunciation, which makes it possible to realize a speech recognition system that is capable of significantly improving a recognition rate as a whole even in the case of large vocabulary. [0071]

Still further, according to the present invention, when there exist words whose pronunciation given by an already-existing database contains "EI" in roman letter notation, words are added by changing "EI" into "EE" and their use frequencies are set based on a prescribed ratio, so it becomes possible to recognize the words even when they are vocalized in manners different from the pronunciation, which makes it possible to realize a speech recognition system that is capable of significantly improving a recognition rate as a whole even in the case of large vocabulary. [0072]

Also, according to the present invention, words are added by long-sounding vowel sounds in pronunciation given by an already-existing database and their use frequencies are set based on a prescribed ratio, so recognition becomes possible even when vocalization is performed in manners different from the pronunciation, which makes it possible to realize a speech recognition system that is capable of significantly improving a recognition rate as a whole even in the case of large vocabulary.

[0073]

Also, according to the present invention, words are added by inserting pauses between syllables of pronunciation given by an already-existing database and their use frequencies are set based on a prescribed ratio, so recognition becomes possible even when vocalization is performed in manners different from the pronunciation, which makes it possible to realize a speech recognition system that is capable of significantly improving a recognition rate as a whole even in the case of large vocabulary.

[0074]

Also, according to the present invention, when there exist words whose pronunciation given by an already-existing database contains double consonants, words are added by changing the double consonants into "TSU" and their use frequencies are set based on a prescribed ratio, so recognition becomes possible even when vocalization is performed in manners different from the pronunciation, which makes it possible to realize a speech recognition system that is capable of significantly improving a recognition rate as a whole even in the case of large vocabulary.

[0075]

Also, according to the present invention, the contents of a database are classified in accordance with a prescribed criterion, use frequencies are estimated, speaker matching is performed at the time of recognition, and speaker matching scores are added to

use frequency scores and acoustic scores at a certain ratio, so it becomes possible to obtain favorable recognition performance, which makes it possible to realize a speech recognition system that is capable of significantly improving a recognition rate as a whole even in the case of large vocabulary.

[Brief Description of the Drawings]

[FIG. 1] A block diagram showing a construction of a first embodiment of the speech recognition system according to the present invention

[FIG. 2] A timing chart for explanation of an operation of a second embodiment of the speech recognition system according to the present invention

[FIG. 3] A timing chart for explanation of an operation of a third embodiment of the speech recognition system according to the present invention

[FIG. 4] A block diagram showing a construction of a fourth embodiment of the speech recognition system according to the present invention

[FIG. 5] A flowchart showing an operation of a use frequency estimation section in a fifth embodiment of the speech recognition system according to the present invention

[FIG. 6] A table for explaining the result of processing by a use frequency estimation section in the fourth embodiment of the speech recognition system according to the present invention

[FIG. 7] A table for explaining the result of processing by the use frequency estimation section in the fifth embodiment of the

speech recognition system according to the present invention

[FIG. 8] A block diagram showing a construction of a sixth embodiment of the speech recognition system according to the present invention

[FIG. 9] A block diagram showing a construction of a conventional speech recognition system

[FIG. 10] A timing chart for explaining an operation of a voice section detection section of the speech recognition system shown in FIG. 9

[FIG. 11] A characteristic curve diagram for explanation of recognition performance of the conventional speech recognition system

[Description of Reference Numerals]

1 acoustic analysis section

2 voice section detection section

3 collating section

4 result output section

5 word dictionary storage section

6 word dictionary sorting section

7 word-dictionary-with-frequency storage section

10 word dictionary storage section

11 use frequency score calculation section

12 database

13 use frequency estimation section

14 speaker recognition section

FIG. 1

1 ACOUSTIC ANALYSIS SECTION

2 VOICE SECTION DETECTION SECTION

3 COLLATING SECTION

4 RESULT OUTPUT SECTION

7 WORD-DICTIONARY-WITH-FREQUENCY STORAGE SECTION

10 WORD DICTIONARY STORAGE SECTION

11 USE FREQUENCY SCORE CALCULATION SECTION

FIGS. 2, 3 & 10

THRESHOLD VALUE P1

VOICE SIGNAL POWER

COMPUTATION BY COLLATING SECTION

COMPUTATION BY RESULT OUTPUT SECTION

TIME

FIG. 4

1 ACOUSTIC ANALYSIS SECTION

2 VOICE SECTION DETECTION SECTION

3 COLLATING SECTION

4 RESULT OUTPUT SECTION

7 WORD-DICTIONARY-WITH-FREQUENCY STORAGE SECTION

10 WORD DICTIONARY STORAGE SECTION

11 USE FREQUENCY SCORE CALCULATION SECTION

12 DATABASE

13 USE FREQUENCY ESTIMATION SECTION

FIG. 5

ST1 IS "OU" CONTAINED?

ST2 ADD WORD FOR "OO"

ST3 IS "EI" CONTAINED?

ST4 ADD WORD FOR "EE"

ST5 IS DOUBLE CONSONANT CONTAINED?

ST6 ADD WORD BY CHANGING DOUBLE CONSONANT INTO "TSU"

ST7 ADD WORD BY LONG-SOUNDING VOWEL SOUND AND NASAL SOUND AND BY INSERTING PAUSE BETWEEN SYLLABLES

FIG. 6

WORD

ABE

SATOU

NITTA

SEINO

USE FREQUENCY

FIG. 7

ABE

A-BE

A·BE

SATOU

SATOO

SA-TO-U

SA-TO-O

SA·TO·U

SA·TO·O

NI-TTA

NITSUTA

NI-TSU-TA

NI·TSU·TA

SEINO

SEENO

SE-I-NO

SE-E-NO

SE·I·NO

SE·E·NO

USE FREQUENCY

FIG. 8

1 ACOUSTIC ANALYSIS SECTION

2 VOICE SECTION DETECTION SECTION

3 COLLATING SECTION

4 RESULT OUTPUT SECTION

7 WORD-DICTIONARY-WITH-FREQUENCY STORAGE SECTION

10 WORD DICTIONARY STORAGE SECTION

11 USE FREQUENCY SCORE CALCULATION SECTION

12 DATABASE

13 USE FREQUENCY ESTIMATION SECTION

14 SPEAKER RECOGNITION SECTION

FIG. 9

1 ACOUSTIC ANALYSIS SECTION

2 VOICE SECTION DETECTION SECTION

3 COLLATING SECTION

4 RESULT OUTPUT SECTION

5 WORD DICTIONARY STORAGE SECTION

6 WORD DICTIONARY SORTING SECTION